Comparaison de mesures de voisment/non-voisement dans les signaux de parole

Stéphane Rossignol, Olivier Pietquin

Comparaison de mesures de voisment/non-voisement dans les signaux de parole

Stéphane ROSSIGNOL, Olivier PIETQUIN

Supélec – Campus de Metz, Équipe IMS
2 rue Édouard Belin, F-57070 Metz, France
Stephane.Rossignol@supelec.fr,Olivier.Pietquin@supelec.fr

Résumé – Cet article présente une mesure de voisement basée sur le calcul du signal analytique. Cette mesure peut être utile pour plusieurs applications concernant le traitement de la parole. Par exemple, considérant la reconnaissance automatique de la parole, elle pourrait être incorporée dans les vecteurs acoustiques communément utilisés, comme par exemple les Mel Frequency Cepstral Coefficients (MFCC) et leurs deux premières dérivées, ceci pour améliorer les performances du système. La base de données TIMIT est segmentée manuellement en phonèmes : c'est pourquoi l'évaluation de la mesure développée est effectuée sur cette base. L'information de voisement est déduite de cette segmentation. Il est montré dans cet article que la segmentation automatique voisé/non-voisé obtenue en utilisant la méthode décrite dans les sections suivantes et la segmentation manuelle voisé/non-voisé fournie dans TIMIT sont très similaires.

Abstract – This paper proposes a voiced/unvoiced measure based on the Analytic Signal computation. This measure can be useful for many speech processing applications. For instance, considering speech recognition, it could be incorporated into commonly used acoustic feature vectors, such as for example the Mel Frequency Cepstral Coefficients (MFCC) and their first two derivatives, in order to improve the performance of the overall system. TIMIT is manually segmented into phones: this is why the evaluation of the developed measure is performed on this database. The voicing information is derived from this segmentation. It is shown in this paper that the automatic voiced/unvoiced segmentation obtained using the method described in the next sections and the manual voiced/unvoiced segmentation provided by TIMIT are very similar.

1 Introduction

Parmi les mesures de voisement/non-voisement de l'état de l'art, l'énergie, le taux de passage par zéro (ou ZCR), l'analyse des coefficients d'autocorrélation, l'analyse du spectre, etc. ([1], [2], [3]) sont très répandues. Elles requièrent de faire l'hypothèse de stationnarité du signal sur quelques millisecondes ou dizaines de millisecondes, et donc requièrent un processus de découpage en trames. Dans cet article, une méthode basée sur le signal analytique (AS) est décrite. Elle utilise des trames très courtes (peu d'articles traitent de l'analyse de trames très courtes). Elle est comparée à deux méthodes simples de l'état de l'art. L'un des objectifs de cet article est de montrer jusqu'à quel point il est possible de réduire la taille T des trames utilisées, tout en gardant de bonnes performances de détection.

2 Description de la méthode

2.1 Synoptique de la méthode

L'analyse du son est effectué en quatre étapes (figure 1).

Premièrement, le signal audio est filtré par un filtre passe-bande, avec une décroissance en $1 - \sqrt{f/f_s}$ dans la bande, et avec $f_1 = 50Hz$ et $f_2 = 300Hz$. Le but de ce filtrage est, dans les parties voisées du son, de rendre le premier partiel aussi prédominent que possible en amplitude ([5], où une telle

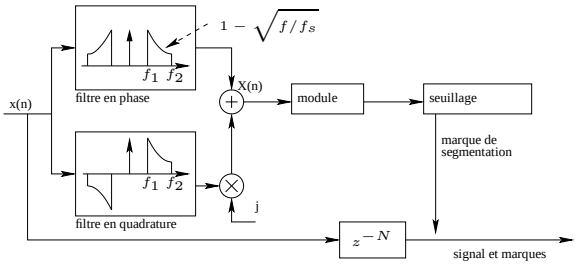

FIG. 1 – Synoptique de la méthode

méthode, basée sur le renforcement du premier partiel pour l'extraction du pitch, est présentée). Donc, le signal $s(n) \simeq a_1 \cos(2\pi f_0 n/f_s + \phi_1)$ est obtenu, où f_0 est la fréquence fondamentale et f_s la fréquence d'échantillonnage . Dans les parties non-voisées du son, un bruit est obtenu.

Deuxièmement, le signal analytique est déterminé en utilisant le filtrage de Hilbert. Dans les parties voisées du son, le signal $X(n) \simeq a_1 \exp(2\pi j f_0 n/f_s + j\phi_1)$ est obtenu (section 2.2). Notons que le seuillage passe-bande et le filtrage de Hilbert sont effectués conjointement.

Troisièmement, le module $A = |X(n)|$ du signal analytique est estimé. Dans les parties voisées du son, il faut remarquer que ce module fournit une estimation approximative de l'amplitude du premier partiel ($A \simeq a_1$). Dans les parties non-voisées du son, seulement du bruit est obtenu. Il est donc fait

l'hypothèse que A a des valeurs plus grandes dans les segments voisés que dans les segments non-voisés du son.

Quatrièmement, le module A est automatiquement seuillé. La segmentation voisé/voisé est obtenue. Le seuillage effectué est décrit en détails dans la section 2.3.

2.2 Signal X obtenu

Dans la bande, plus f_0 est proche de f_1 ou de f_2, moins l'approximation de $X(n)$ faite ci-dessus est exacte, comme le montre la figure 2. Nous avons, à gauche le signal simulé original (10 partiels d'égales amplitudes, $f_0 = 150Hz$), le signal $X(n)$ et le fondamental ; à droite le rapport de l'énergie du signal $X(n)$ et de celle du fondamental en fonction de f_0. Le rapport est proche de 1 pour plus de la moitié de la bande.

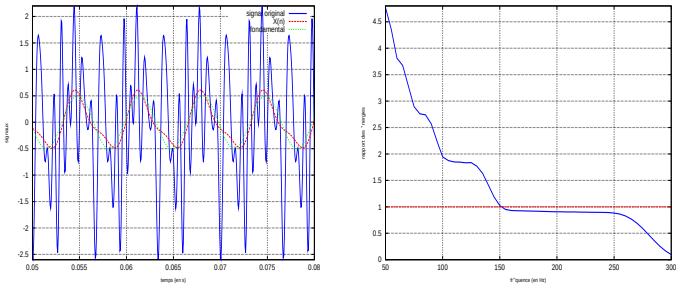

FIG. 2 – Gauche : signal simulé original, signal $X(n)$ et fondamental ; Droite : énergie de $X(n)$ versus f_0

2.3 Le seuillage

Les méthodes de seuillage automatique, venant de la communauté du traitement des images, sont nombreuses (voir [7]). Dans cet article, une méthode « ad-hoc » et la méthode d'Otsu sont testées. Le niveau d'enregistrement des sons analysés (section 3) varie ; de ce fait, les seuils doivent être réestimés pour chacun des sons analysés. Ceci se traduit dans la définition du seuil « ad-hoc » donnée ci-dessous par le fait que la valeur moyenne du signal A à seuiller intervient. La valeur t_3 de ce seuil est égale à $t_3 = C_3 \text{mean}[A]$. C_3 est une constante utilisée pour tous les sons de la base de données. Une étape d'apprentissage est nécessaire pour déterminer la valeur de cette constante. La méthode d'Otsu est expliquée en détails dans [7]. Le seuil optimal t_O est obtenu. Dans [6], app. B, les performances de 19 méthodes de seuillage sont comparées, considérant trois paramètres perturbateurs : 1. les deux classes ne sont pas balancées ; 2. la variance d'une des classes est plus grande que la variance de l'autre ; 3. les densités de probabilité des deux classes se superposent. Considérant la problème de segmentation voisé/non-voisé, les techniques sélectionnées doivent rester robustes avant tout au second de ces paramètres. En effet, la variance de l'amplitude du signal analytique, de l'énergie et du ZCR, au cours des parties voisées du signal, est grande, du fait que ces caractéristiques au cours d'un phonème et d'un phonème à l'autre, peuvant varier énormément ; au contraire, l'amplitude du signal analytique, de l'énergie et du ZCR, au cours des parties non-voisées du signal, varie beaucoup moins au cours d'un segment et d'un segment à l'autre. À cet égard, la méthode d'Otsu a montré dans [6] qu'elle est une des plus robustes méthodes de seuillage. C'est la raison pour laquelle elle est étudiée dans cet article.

3 Expérimentations – quelques résultats

3.1 Base de données

Les évaluations sont effectuées sur la base de donnée TIMIT ([4]). TIMIT contient un total de 6300 phrases, 10 phrases pour chacun des 630 locuteurs utilisant 8 des dialectes les plus importants des États-Unis. TIMIT est segmentée en phonèmes. Ceci conduit à un total de 61 labels de segmentation différents, desquels l'information de voisement est dérivée (ceci sans tenir compte du dévoisement et du fait que certains phonèmes sont voisés ou non-voisés selon le contexte, ce qui introduit dans nos mesures un taux d'erreur ; nous le considérons faible). Le problème de segmentation voisé/non-voisé est plutôt balancé. En effet, environ 56 % des échantillons, parmi plusieurs centaines de millions, sont labelés voisés, et 44 % non-voisés.

3.2 Comportement des seuils

Expérimentalement, il peut être noté que la méthode d'Otsu donne systématiquement des valeurs de seuil surestimées. Pour modifier la méthode originale, une constante multiplicative C_O est ajoutée. Est considéré alors le seuil $t_O^M = t_O C_O$. Une étape d'apprentissage est nécessaire afin de déterminer la valeur de C_O. Dans le tableau 1, les résultats obtenus avec la méthode d'Otsu modifiée sont montrés. Ils sont similaires à ceux obtenus en utilisant le seuil « ad-hoc ».

TAB. 1 – C_3 et C_O apprises et taux de bonne détection (TBD) obtenu sur TIMIT avec le signal analytique (AS), l'énergie (E) et le taux de passage par zéro (ZCR) ; $T = 14\ ms$

	C_3	TBD (%)	C_O	TBD (%)
AS	3.4162e-1	92.599	2.1949e-1	92.089
E	4.7992e-2	81.767	1.2366e-2	81.416
ZCR	9.9290e-1	89.066	1.1212e0	88.476

Du fait que les meilleures performances sont obtenues pour une trame de 14 ms (voir figure 3) pour les méthodes AS et E et du fait que les performances de la méthode ZCR pour cette taille de trame sont proches des meilleures performances obtenues en utilisant cette méthode, il a été décidé de ne montrer dans cette section que les résultats d'apprentissage pour les constantes des seuils obtenus considérant cette taille de trame. Quand la méthode AS est utilisée, un taux de bonne détec-

tion compris entre 92.089 % (avec la méthode d'Otsu modifée) et 92.599 % (avec le seuil « ad-hoc ») est obtenu. Quand l'énergie est utilisée un taux de bonne détection compris entre 81.416 % (avec la méthode d'Otsu modifiée) et 81.767 % (avec le seuil « ad-hoc ») est obtenu. Quand le ZCR est utilisé un taux de bonne détection compris entre 88.476 % (avec la méthode d'Otsu modifiée) et 89.066 % (avec le seuil « ad-hoc ») est obtenu. La méthode d'Otsu originale donne un taux de bonne détection de seulement 74.602 % (avec la méthode AS) et de 54.924 % (avec l'énergie). Avec le ZCR, le taux de bonne détection est moins détérioré, puisqu'il vaut encore 88.409 %. Dans la suite de l'article, le seuil « ad-hoc » est retenu, pour les trois mesures de voisement comparées ici.

3.3 Performances selon la longueur des trames

La figure 3 montre que la caractéristique basée sur l'AS développée dans cet article fournit des résultats bien meilleurs que les caractéristiques basées sur l'énergie et le ZCR.

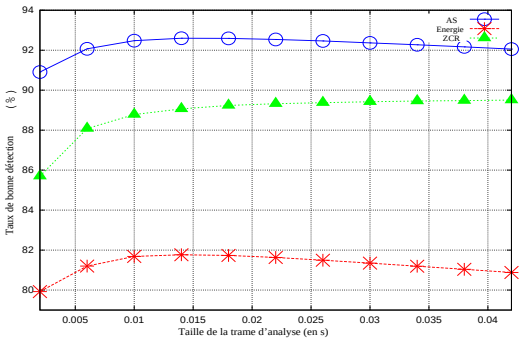

FIG. 3 – TBD obtenu pour différentes longueurs de trames

3.4 Aspects temporels – précision dans la position des marques de segmentation trouvées

Dans cette section, la précision dans la position des ruptures trouvées entre un segment voisé (respectivement non-voisé) et un segment non-voisé (respectivement voisé), comparées aux positions fournies dans TIMIT, est étudiée. Une marque de segmentation est positionnée sur l'échantillon (i) si pour cet échantillon il fut décidé dans les précédentes étapes de l'analyse que le signal est voisé (respectivement non-voisé) et si pour le précédent échantillon ($i-1$) il fut décidé que le signal était non-voisé (respectivement voisé).

Il faut remarquer que, dans la parole naturelle, des successions de plusieurs phonèmes voisés ou de plusieurs phonèmes non-voisés sont possibles. Les méthodes utilisées dans cet article ne peuvent pas déterminer les frontières entre des phonèmes voisés successifs ou entre des phonèmes non-voisés successifs. De ce fait, il n'est pas possible d'allouer à chaque marque de segmentation TIMIT une des marques de segmentation trouvées. Au lieu de cela, une des marques de segmentation TIMIT est allouée à chaque marque de segmentation trouvée. La marque de segmentation TIMIT la plus proche est sélectionnée. La distance d en ms entre une marque de segmentation trouvée et la marque de segmentation TIMIT correspondante caractérise la précision dans la position des marques de segmentation trouvées. Cependant, des fausses détections peuvent survenir. C'est-à-dire : une transition est détectée au cours d'un segment voisé ou au cours d'un segment non-voisé. Dans ce cas, la distance d est trop grande et l'allocation doit être rejetée. Il est fait l'hypothèse dans cet article que le taux de parole en terme de phonèmes par seconde en anglais est d'environ 12, conduisant à une longueur d'environ 80 ms pour chaque phonème. Il est donc décidé dans cet article que si d est supérieur à 30 % de cette durée, l'allocation doit être rejetée.

La figure 4 montre que, pour les trames les plus longues, comme attendu, d diminue comme la longueur de trame diminue. Cependant, pour les trames plus courtes, d augmente comme la longueur de trame décroît. Ceci est dû au fait que pour les trames les plus courtes, l'analyse devient moins stable, donnant de rapides successions de marques de segmentation, c'est-à-dire un très court segment voisé (respectivement non-voisé) suivi par un très court segment non-voisé (respectivement voisé). Plusieurs des marques de segmentation trouvées ont de ce fait la même marque de segmentation TIMIT en tant que marque allouée, conduisant à une imprécision plus grande.

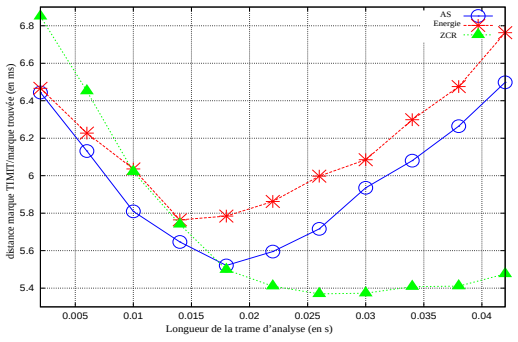

FIG. 4 – Précision dans la position temporelle des marques de segmentation trouvées pour différentes longueurs de trames

3.5 Robustesse au bruit

Dans cette section, la robustesse au bruit de la méthode proposée est démontrée. Un bruit normal est artificiellement ajouté aux sons de TIMIT, de telle façon que le rapport signal à bruit (SNR) obtenu est le même pour chacun d'eux. Pour une trame de 14 ms, les résultats obtenus sont montrés sur la figure 5. Pour les petits SNR, les méthodes AS et E restent robustes. Le taux de bonne détection ne décroît que peu, quand le SNR est au-dessous de 9 dB. À 5 dB, le taux de bonne détection, considérant la méthode AS, vaut encore 89.718 %, et considérant l'énergie il vaut encore 81.065 %. Au contraire, pour la méthode ZCR, le taux de bonne détection décroît à 79.603 % pour un SNR de 9 dB et à 74.023 % pour un SNR de 5 dB.

De plus, il peut être remarqué sur la figure 6 (gauche) qu'en ce qui concerne la méthode AS, la valeur de C_3 obtenue après apprentissage ne dépend quasi pas du SNR. Au contraire, la valeur de C_3 obtenue après apprentissage pour l'énergie est plus que triplée quand le SNR vaut 9 dB, et plus que quintuplée quand le SNR vaut 5 dB. Ceci indique que la méthode AS ne requiert pas absolument qu'on ait avant de l'appliquer une estimation du SNR, ce qui n'est définivement pas le cas pour l'énergie. La figure 6 (droite) montre ceci. C_3 est pris constant et égal à sa valeur quand il n'y a pas de bruit additif. On peut voir qu'en ce qui concerne l'énergie, le taux de bonne détection décroît de 79.266 % à 55.208 % quand le SNR décroît seulement de 13 dB à 10 dB, et que ceci est bien moins le cas en ce qui concerne la méthode AS.

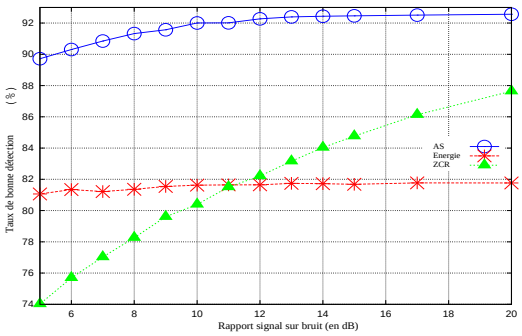

FIG. 5 – TBD, pour différents SNR ; $T = 14\ ms$

La valeur de C_3, en ce qui concerne le ZCR, ne dépend pas du SNR. Ceci montre simplement le fait que cette méthode est moins robuste au bruit que les deux autres. Cependant, si C_3 est pris constant et égal à sa valeur quand il n'y a pas de bruit additif, le ZCR montre un comportement plutôt robuste, et se révèle presque compétitif avec la méthode AS quand le SNR est très bas. Il faut noter que des résultats similaires sont obtenus quand des trames plus grandes sont utilisées.

4 Conclusion

Dans cet article, plusieurs méthodes pour automatiquement effectuer la segmentation voisé/non-voisé des signaux de parole monophoniques sont comparées. Il est montré que la méthode développée dans cet article, basée sur le signal analytique, fournit les résultats les plus fiables. La meilleure information de voisement est obtenue en utilisant des trames longues de 14 ms, puisqu'alors un taux de bonne détection de 92,599 % est atteint. De façon similaire, une grande précision en terme de localisation temporelle des marques de segmentation est obtenue. Les meilleurs résultats sont obtenus pour une longueur de trame de 18 ms. Il faut remarquer que dans ce cas, notre mesure de voisement reste fiable ; le taux de bonne détection est encore égal à 92,593 %. De plus, notre méthode est beaucoup plus robuste au bruit que les deux autres. Pour une trame de 14 ms, avec un SNR de 5 dB, le taux de bonne détection est d'environ 90 % avec la méthode AS, alors qu'il est de 81 % avec l'énergie et de 74 % avec le ZCR.

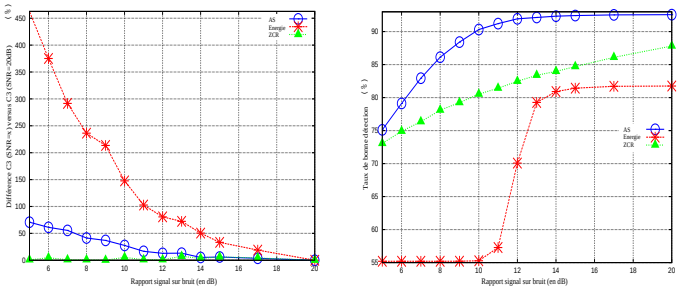

FIG. 6 – Gauche : Différence en % de la valeur de C_3 apprise pour différents SNR, la valeur obtenue pour un SNR de 20 dB étant prise comme référence ; Droite : TBD, pour différents SNR, C_3 pas ajusté au SNR ; $T = 14\ ms$

Des améliorations de la technique sont envisagées. La structure périodique des signaux voisés n'est pas prise en compte pour le moment. La méthode AS peut fournir une estimation du pitch f_0, et donc peut être adaptée dans le but de prendre avantage de ceci. De plus, doivent être effectués : l'étude de la robustesse de la méthode à des bruits autres que blancs, notamment dans une bande passante basse (milieu de type habitacle de voiture) ; des comparaisons avec d'autres mesures de voisement ; maintenant que la méthode est validée sur TIMIT des tests avec signaux plus compliqués que ces sons. Utiliser l'information précise de voisement/non-voisement obtenue avec la méthode décrite dans cet article, par exemple dans les systèmes existants de reconnaissance de la parole et les systèmes d'alignement de la parole, peut améliorer leurs performances et peut améliorer leur rapidité.

Références

[1] B. S. Atal et L. R. Rabiner. A Pattern Recognition Approach to Voiced – Unvoiced – Silence Classification with Applications to Speech Recognition. IEEE Transaction on Acoustics, Speech, and Signal Processing, 1976.

[2] M. Greenwood et A. Kinghorn. SUVing : Automatic Silence/Voiced/Unvoiced Classification of Speech. Undergraduate Coursework – Department of Computer Science, University of Sheffield, UK, 1999.

[3] S. Rossignol, X. Rodet, J. Soumagne, J.-L. Collette et P. Depalle. Feature Extraction et Temporal Segmentation of Acoustic Signals. Proceedings of the ICMC, 1998.

[4] Linguistic Data Consortium. DARPA TIMIT Acoustic-Phonetic Continuous Speech Corpus. NIST Speech CD 1-1.1, 1990.

[5] W. Hess. Pitch Determination of Speech Signals. Springer-Verlag, 1983.

[6] S. Rossignol. Segmentation et indexation de signaux sonores musicaux. 2000. Université de Paris VI.

[7] P. K. Sahoo, S. Soltani et K. C. Wong. A Survey of Thresholding Techniques. Computer Vision, Graphics, and Image Processing. 1988.